尺度无关的级联卷积神经网络人脸检测算法 *

郑成浩[a]，刘　兵[a, b]，周　勇[a]

(中国矿业大学 a. 计算机学院; b. 中国科学院电子研究所, 江苏 徐州 221116)

摘　要：卷积神经网络在进行图片处理时需要输入固定尺寸大小的图片，该限制会导致原图在放缩过程中损失大部分信息。另外，目前人脸检测算法多用单一结构网络进行特征提取，这就使得算法的泛化能力较弱。针对以上两个问题，提出了一种将级联卷积神经网络与空间金字塔池化相结合的人脸检测算法。该方法将三级卷积神经网络模型连接起来，其中三级神经网络模型之间各不相同，结构从简单到复杂，在不同层次的神经网络上提取不同的人脸特征并筛选图片，完成对图片中人脸区域的检测。同时，在每级网络层次中加入空间金字塔池化层，这种池化策略无须固定尺寸大小的输入，增加了模型输入的尺寸选择。在标准人脸数据集中，该方法相对于传统方法实现了模型的多尺度输入，提升了检测的性能，并降低了检测人脸的时间。

关键词：级联卷积神经网络；空间金字塔池化；人脸检测

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2017.08.0957

Face detection algorithm based on scale-independent cascade convolution neural network

Zheng Chenghao[a], Liu Bing[a, b], Zhou Yong[a]

(*a. College of Computer Science & Technology*, *b. Institute of Electrics*, *Chinese Academy of Sciences*, *China University of Mining & Technology*, *Xuzhou Jiangsu* 221116, *China*)

Abstract: Since the convolution neural network needs to input a fixed size image when performing image processing, this will lead to the loss of most of the original information in the scaling process. In addition, the feature extraction of images will not be put in place when the network has only one structure. To solve the above two problems, this paper presented a face detection algorithm combining cascade convolution neural network and spatial pyramid pooling. In this method, it cascaded three different convolution neural network models, the structure of which were from simple to complex, and extracted different face features at different levels to complete the detection of the face areas of images. At the same time, it added the pyramid pool at each level of the network, and this pooling strategy did not require a fixed size input, increasing dimension selection of model input. Compared with the traditional method, this method realizes the multi-scale input of the model, improves the detection performance, and reduces the time of face detection in the standard face data set.

Key Words: cascade convolution neural network; spatial pyramid pooling; face detection

0　引言

人脸检测是目标检测识别中一个热点研究领域，即采用一定的策略在任意一幅给定的图片中检测其中的人脸区域，以返回人脸的位置或者其他信息。在早期的图像识别系统中，主要是通过尺度不变特征变换(scale-invariant feature transform,SIFT)和方向梯度直方图(histogram of oriented gradients,HOG)等方法进行特征提取，然后将提取到的特征输入分类器中进行图像识别。上述方法所得到的特征本质上是人工设计的特征，对于不同的检测识别问题，所提取的特征对系统的性能有着显著地影响，因此这需要研究人员对所要解决的问题有着非常深入的了解，这样才有可能设计出对解决该问题效果较好的特征，从而提升系统的性能。这个时期的图像检测识别系统大多是针对单一特定问题的解决方案，系统整体的泛化能力较差；另外，当时系统的所处理的数据量规模都较小，难以在实际问题中实现准确的识别效果[1]。

深度学习是机器学习的一个分支，是近些年来机器学习领域取得的重大突破和研究热点之一[2~6]。2011 年以来，研究人员首先在语音识别问题上应用深度学习技术，将识别的准确率提高了 20%～30%，取得了十年来最具突破性的进展。仅仅一年之后，基于卷积神经网络的深度学习模型就在大规模图像识别分类任务上取得了非常大的性能提高，掀起了深度学习的热

基金项目：国家自然科学基金青年科学基金资助项目（61403394）；国家自然科学基金面上项目（61572505）

作者简介：郑成浩（1992-），男，江苏徐州，硕士研究生，主要研究方向为机器学习、模式识别（08133323@cumt.edu.cn）；刘兵（1981-），男，副教授，硕导，主要研究方向为机器学习、模式识别；周勇（1974-），男，院长，教授，博导，主要研究方向为数据挖掘、无线传感器网络.

潮[7,8]。

但是在卷积神经网络进行图像处理时，首先要确保的就是输入神经网络模型的图片大小需要保持一致。原因在于当卷积神经网络进行卷积，池化等一系列操作之后将会把得到的数据输出到最后的全连接层中，而全连接层的神经元数量是固定的，这也就意味着与全连接层相连的权值数量要保持固定，如果权值数量发生改变，则无法进行权值的计算或者更新。因此，几乎所有卷积神经网络在进行图片处理之前都需要将输入图片的大小放缩或者裁剪成统一尺寸，这样才能进行卷积神经网络的训练或者测试。

针对这一问题， He 等人[9]提出了可以实现卷积神经网络多尺度输入的空间金字塔池化算法(spatial pyramid pooling)。空间金字塔池化算法通过在卷积神经网络的全连接层之前加入金字塔池化层，将卷积池化得到的不同尺寸大小的特征图，经过金字塔池化处理之后，形成统一维度大小的数据输出到之后的全连接层中。这样就使得输入卷积神经网络的图片大小不再需要统一的大小，可以让图片本身在预处理阶段有更多的信息被保留下来，在之后用卷积神经网络提取到关键特征的可能性更高。

另外，在许多卷积神经网络进行的图像处理中，更多的是设计一个单一的卷积网络模型，这样神经网络采集的特征也相对单一，使得网络模型的泛化性能不强，往往针对某些问题有较好的效果，在针对其他问题时可能效果一般。

因此，Li 等人[10]在 CVPR2015 上提出了级联卷积神经网络(convolutional neural network cascade) 的算法模型，模型结构简化图如图 1 所示。该模型通过将不同结构的卷积神经网络模型连接起来，逐步地进行特征提取。模型结构由简单到复杂，开始卷积神经网络较为简单，相当于特征的粗提取过程，大致地对输入图片进行初步分类，将分类结果为检测目标的图片输入到下一级卷积神经网络模型中，该级模型相较于上一级则会更加复杂，图片特征的提取过程也会更加细致，再经过这一级图片分类将需要的结果输入到最后的最为复杂的网络中，进行最后的特征提取，最终得到分类结果。该模型利用不同模型提取特征，逐步精确，在降低了检测时间的同时，也提高了准确率。

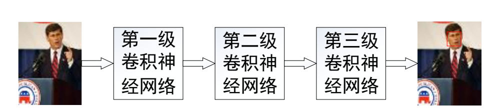


图 1 级联卷积神经网络结构简图

本文通过将金字塔池化与级联卷积神经网络结合，提出了一种尺度无关的深度卷积神经网络人脸检测算法。首先根据级联卷积神经网络的思想设计出三级级联的卷积神经网络模型，然后将金字塔池化方法嵌入到每一级的卷积神经网络中，这样每一级的网络模型都不会受到输入图片尺度大小不同的影响。本文整体人脸检测过程主要分为三步：a)通过滑动窗口扫描待检测图片，得出要检测的所有候选框；b)将候选框输入到训练好的级联神经网络模型中进行人脸图片分类；c)将分类得到人脸图片通过非最大值抑制(NMS)进行最后整合，在原图中标记出人脸区域位置。经过在标准人脸数据集上的实验，本文所提出的算法在检测性能和所用时间上比传统方法都有较大提升。

1 卷积神经网络

卷积神经网络起源于 20 世纪 60 年代初期，Hubel 和 Wiesel 等人通过对猫的大脑视觉皮层系统的研究，提出了感受野的概念，并进一步发现了视觉皮层通路中对于信息的分层处理机制，由此获得了诺贝尔生理学或医学奖。到了 80 年代中期，Fukushima 等人基于感受野概念提出的神经认知机，可以看做是卷积神经网络(convolution neural networks,CNNs)的第一次实现，也是第一个基于神经元之间的局部连接性和层次结构组织的人工神经网络。1990 年，LeCun 等人在研究手写数字识别问题时，首先提出了使用梯度反向传播算法训练的卷积神经网络模型，并在 MNIST 手写数字数据集上表现出了相对于当时其他方法更好的性能。目前，卷积神经网络已成为当前图像识别领域的研究热点，它是第一个真正意义上的成功训练多层神经网络的学习算法模型，对于网络的输入是多维信号时具有更明显的优势[11~13]。

1.1 级联卷积神经网络

由于最近几年在人脸检测领域方面的研究主要集中在所识别的人脸区域中不可控部分的问题上，如夸张的表情、姿势改变、面部遮挡等[14]。这些情况都会影响到人脸检测的最终效果。另外，面对如此多的问题，仅靠单一结构的模型进行检测很难产生良好的泛化能力，使得模型在实际应用中的鲁棒性较低。因此，现在人脸检测方面的三个主要难点是：a)人脸在杂乱情景中的可变情况太多；b)图片中可能的人脸位置和人脸大小的数量太多；c)单一结构的模型在情况多变的问题上鲁棒性不强。

针对以上问题， Li 等人在文献[10]中提出的级联卷积神经网络模型，有效地解决了上述面临的主要难点。级联卷积神经网络的理论来源主要是 2001 年 Viola 等人[15]提出的简单特征的级联加速器(boosted cascade of simple features)的算法。该算法建立了一种将简单特征集合起来作为分类器的思路。虽然在这之后提出了很多 V-J 算法进行改进的方案，但在改进的过程中所设定的级联数目以及所使用的简单特征都会影响到最终的检测精度，而且所选取的特征往往需要人工设计，在特征选取出现误差时，整个模型的检测效果就会降低很多，同时对复杂情况的泛化性能也不强。正是由于这样的原因，Li 等人选择了卷积神经网络作为特征提取的方式。与之前人工所选取的特征不同，卷积神经网络可以自动地去学习特征，并且可以在训练大量样本的过程中捕捉到人脸区域中各种复杂多变的情况，这对于构建一个能够精确获取人脸特征的人脸检测算法是极为重要的。

Li 等人提出级联卷积神经网络的主要结构如图 2 所示。从图中可以看出，该网络模型主要由三级卷积神经网络组成。其中每级包括一个二分类网络(12-net, 24-net, 48-net)和一个校准

网络(12-calibration,24-calibration, 48-calibration)。这三级网络首先在输入图片的分辨率上有所不同，分辨率逐渐加大主要是为了逐渐提高识别精度，而且可以减少运行时间，提高模型效率；其次三级网络的结构各不相同，可以明显地看出网络结构从简单到复杂，前面简单的网络进行特征的粗提取，后面复杂的网络将对前面筛选出来的图片进行更加精确的分类。整个网路的工作流程就是首先将待检测图片输入第一级二分类网络中进行特征分类，如果第一级二分类网络判定是人脸图片，则将会把图片输入到第一级矫正网络中进行位置矫正，如果不是人脸图片，则会直接剔除这张图片进行下一张图片的判断；然后将判定为人脸的图片输入到第二级二分类网络中，再进行与之前相同的操作；经过最后一级网络得到的人脸区域图片再通过非最大值抑制(non-maximum suppression,NMS)筛选，最终在原图中标记出人脸位置。

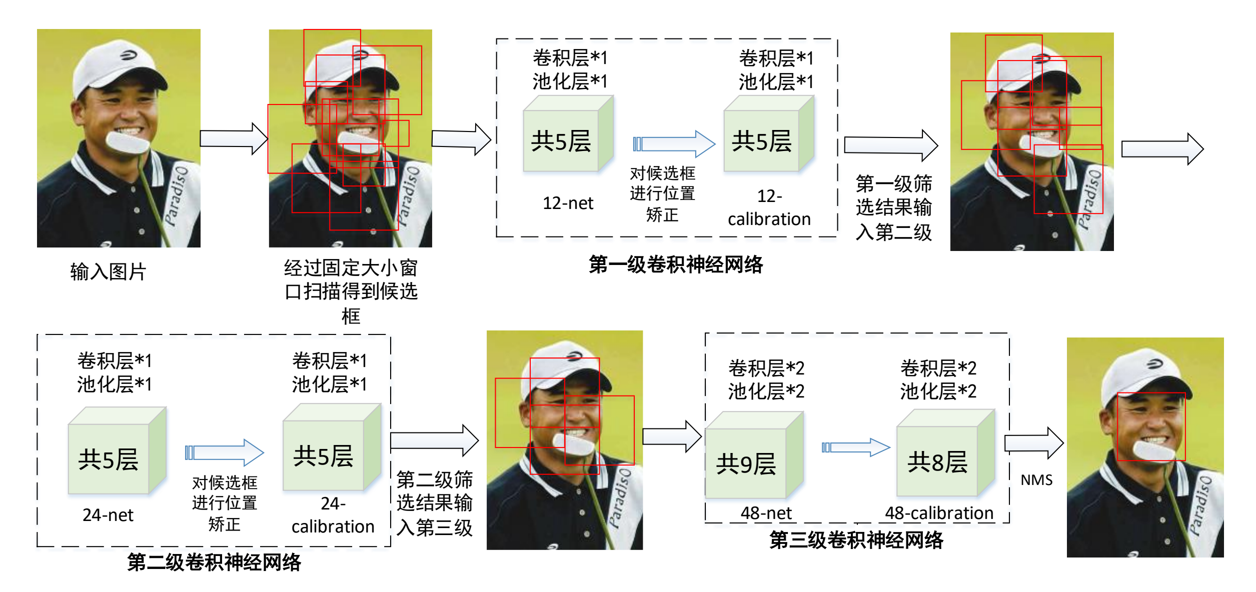


图 2　级联卷积神经网络主要结构

该模型在标准人脸数据集的测试中有着不错的准确率。另外，得益于前两级较为简单的网络结构，使得模型整体的检测速率也有所提高，所用时间相对于其他传统网络明显减少。

1.2　空间金字塔池化

1.2.1 空间金字塔池化介绍

对于现在多数在使用的卷积神经网络而言，都要求卷积神经网络的输入尺寸是固定大小的，这就要求卷积神经网络在训练或者测试之前需要将数据的输入尺寸放缩到相同的尺寸大小。例如著名的卷积神经网络模型 AlexNet，在进行图像处理时要求输入图片的尺寸统一为 227×227。也就是说，将输入数据放缩成统一的大小尺寸是训练或者测试卷积神经网络的首要步骤。但是在数据预处理时，比如将原始图片放缩或者裁剪成统一大小(图 3)，当输入图片的尺度发生变化时，传统卷积神经网络将无法实现图片的多尺度输入,同时在统一大小的过程相对于多尺度的预处理过程会有更多数据的损失，这对之后的训练和测试都会有一定的影响。

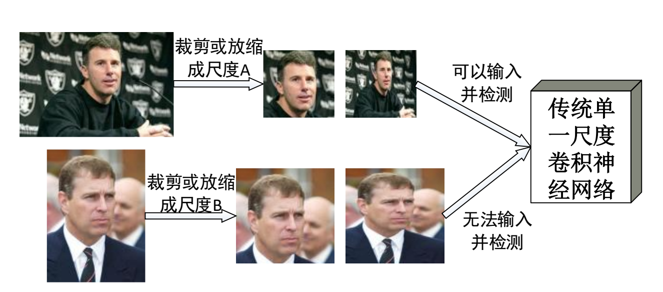


图 3　图片预处理

针对这一问题，He 等人在文献[9]中提出了空间金字塔池化(spatial pyramid pooling,SPP)的算法来解决输入数据的尺度变化问题。由于卷积层和池化层都不要求固定尺寸大小的输入，只有在经过卷积层和池化层之后的全连接层需要固定大小的输入。这里可以假设输入图片的大小是 100×100，经过 5 个 3×3 的卷积核之后会产生 5×98×98 的特征图，而当输入图片的大小变成 102×102 时，在经过相同操作之后就会得到 5×100×100 的特征图，那么两种大小的输入在经过 2×2 的池化之后会分别得到 25×25 和 26×26 的特征图。因此，从这里可以看出，卷积层和池化层可以处理任意输入大小的图片，无须进行调整输入尺寸的预处理操作。但是在之后的全连接层是需要对输入的大小有要求的。假设最后一个卷积层有 50 个输出，下一层全连接层有 1 000 个神经元，那么这个连接矩阵的维度就是 50×1000；而如果网络每次的输入图片大小不一样，那么之后与全连接层连接的矩阵维度将会发生改变，使得网络无法训练或者测试。所以，空间金字塔池化算法就是在全连接层之前加入了一个空间金字塔池化层，以保证输入到全连接层的任意大小的图片都被处理成相同维度的数据。引入金字塔池化的卷积神经网络如图 4 所示。从图中可以明显看出，输入图像在引入金字塔池化后无须进行预处理，实现了卷积神经网络的多尺度输入。

1.2.2 空间金字塔池化具体算法

如图 5 所示，从上往下看，这是一个传统的网络架构模型，卷积层后面连接着全连接层。这里需要处理的就是在网络的全连接层之前加入金字塔池化层来解决输入图片大小不一的情况。

可以看出这里的金字塔池化层就是把前一卷积层得出的特征图进行 3 个池化操作：最上边池化操作是对原始特征图进行池化，中间是把特征图分成 4 份进行池化，最下面是把特征图分成 16 份进行池化。这样最终就会连接成一个 16+4+1=21 的特征向量输入到全连接层中。这样就解决了输入图片大小不一致的问题。

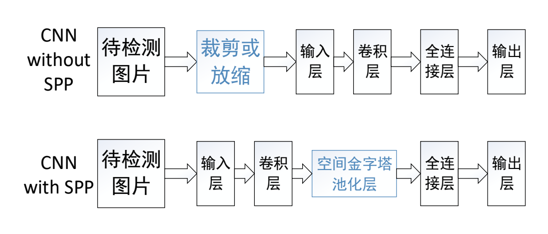


图 4　空间金字塔池化位置图

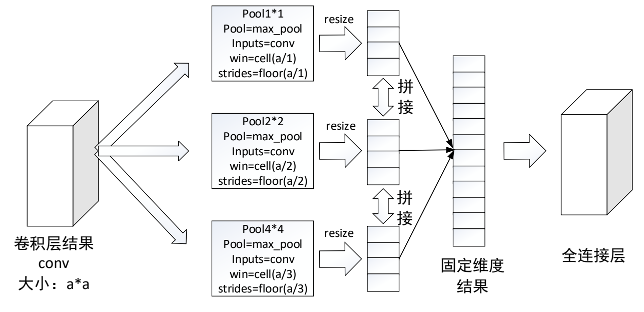


图 5　金字塔池化算法演示

在图片输入大小不等的情况下，假设经过卷积池化操作之后得到的特征图的大小是 a×a，而金字塔池化的通道数是 n，那么每一个窗口的边长是 win=cell(a/n)，池化移动步长是 strides=floor(a/n)，其中 cell 是向上 取整，floor 是向下取整。最终经过金字塔池化会形成如图 5 所示的 n 个池化操作，这些池化操作都会采用基本的池化方法(如最大池化)，只不过使用了不同的窗口大小和移动步长而已。

算法 1　SPP 方法分析过程

输入：卷积池化后的特征矩阵 $X \in \mathbb{R}^{a \times b}$

输出：空间金字塔池化后的特征向量 F

Step1：计算 w = cell(a/n)，h = cell(b/n)，stride1 = floor(a/n)，stride2 = floor(b/n)，n = 1,2,3…；

Step2：用求得的参数分别对特征矩阵 X 进行池化，得到特征 f1,f2,f3…；

Step3：将求得的特征逐个连接，得到新特征 F。

2　基于金字塔池化的级联卷积神经网络模型及算法

2.1　模型设计

虽然级联卷积神经网络在人脸检测领域有着出色的性能，但是算法本身并不支持图片的多尺度输入，导致模型在预处理阶段会损失图片的大部分信息；而空间金字塔池化算法解决了卷积神经网络多尺度输入的问题。因此，本文在级联卷积神经网络的基础上，结合空间金字塔池化算法的优点，提出了一种尺度无关的级联卷积神经网络的人脸检测算法。

算法的模型结构如图 6 所示。在卷积神经网络的全连接层之前加入了金字塔池化层，金字塔池化的通道数统一是 5，这样每一级卷积神经网络就都支持图像的多尺度输入，而且整体网络的结构复杂度是由简单到复杂。与文献[10]相比，本文所提出的算法也采用了三级级联卷积网络结构。为了更快地进行图片检测，本文所提出的算法并没有设置校准网络。这样整体模型仅仅只有三个卷积神经网络，在训练和检测速度上都会大大加快。

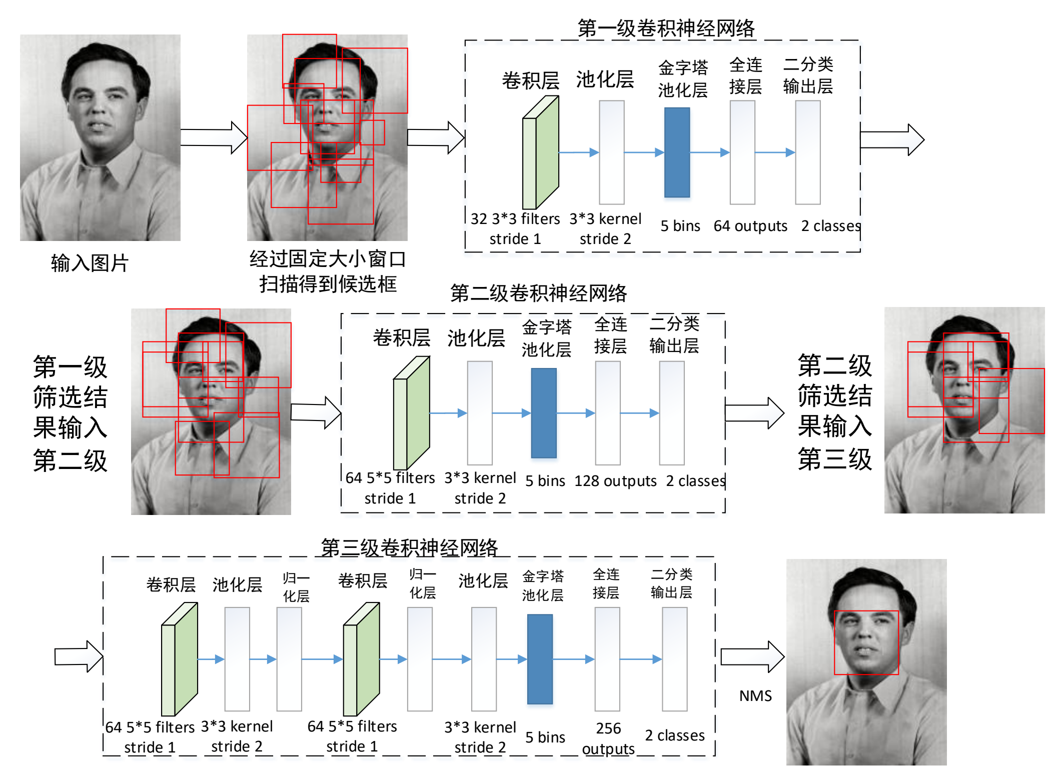


图 6　实验模型结构

2.2　算法描述

1）训练阶段

在训练阶段使用的是 AFLW 人脸数据集作为正样本的来源。在 AFLW 数据集中包含约 2.1 万张图片(基本都是高清图片)，其中标记了约 2.4 万个人脸的矩形框、三维旋转角度、是否遮挡、是否戴眼镜等信息。因此，在进行生成正样本时，按照 AFLW 数据集中给出的标注信息将人脸区域截出，然后对截出的样本进行随机的平移、旋转和翻折操作，最终得出所有的训练正样本。负样本则是从 COCO 数据集中选取不包含人脸的图片进行随机切割，最终得到所有的负样本。正样本的数量大约是 3.5 万张，负样本约 3 万张。

在训练过程中，由于金字塔图像可以解决输入图像多尺度的问题，所以训练时采用了两种分辨率进行训练。首先是将样本统一放缩到 24×24 大小进行训练，在训练结果收敛之后，再将训练好的模型用 12×12 大小尺寸的图片进行训练，当模型再次收敛之后则认为训练过程完成。

2）测试阶段

在测试阶段，需要将待检测图片进行预处理，首先需要将图片进行金字塔图像处理，通过这种方法将图片放缩成包含不同大小的图片的图片组。然后对图片组中的所有图片按照滑动窗口机制(sliding windows)进行操作，这里的滑动窗口将会首先按照 24×24 大小进行操作，产生所有的大小为 24×24 的候选框，并标记好所有候选框在原图中的位置信息，将所有的候选框放入训练好的神经网络模型中。模型这时会在第一级删除候选框部分非人脸图片，保留候选框中的人脸图片，之后的网络依次执行这种操作，直到最后一层得到最终的候选框。最后将候选框经过非极大值抑制(NMS)删除重复性较大的框，得出最终人脸区域的候选框，接着根据候选框的位置信息在原图中标记出人脸区域。之后，再通过 12×12 的滑动窗口得出 12×12 大小的候选框进行对比测试，完善多尺度输入的测试。整体算法流程如算法 2 所示。

算法 2　基于金子塔池化的级联卷积神经网络检测算法

输入：待检测图片

输出：标记出人脸区域的待检测图片

a)利用训练样本进行模型训练；

b)将待检测图片进行金字塔图片处理，得到不同尺度大小的图片组；

c)对图片组中的所有图片进行 24×24(或 12×12)的滑动窗口操作，并记位置信息；

d)将得到的所有候选窗口放入训练完成的第一级卷积神经网络，其中第一级网络包含 1 个卷积层(32 个 3×3 的卷积核)，1 个池化层(3×3 的最大池化)，1 个金字塔池化层(5 个通道)，1 个全连接层(64 个神经元)；

e)将第一级筛选得到的候选窗口放入第二级卷积神经网络，其中第二级网络包含 1 个卷积层(64 个 5×5 的卷积核)，1 个池化层(3×3 的最大池化)，1 个金字塔池化层(5 个通道)，1 个全连接层(128 个神经元)；

f)将第二级筛选得到的候选窗口放入第三级卷积神经网络，其中第二级网络包含 2 个卷积层(64 个 5×5 的卷积核)，2 个池化层(3×3 的最大池化)，2 个归一化层，1 个金字塔池化层(5 个通道)，1 个全连接层(256 个神经元)；

g)将最后得到的候选窗口经过 NMS，去除相似度大于 0.5 的窗口；

h)根据所得候选窗口的位置信息，在原图中标记出检测得到的人脸区域。

3　实验结果分析

3.1　训练阶段实验结果分析

在训练阶段，首先对比了模型中有空间金字塔池化层的情况下，当输入的训练图片尺寸为 12×12 与 24×24 时，模型的训练收敛情况与训练速度的对比；然后进行在输入图片都是 24×24 情况下模型中有无空间金字塔池化层的训练收敛情况与训练速度对比。图 7 中所展示都是第二级卷积神经网络的训练收敛图。

当输入图片大小分别为 12×12 和 24×24 时，训练收敛情况如图 7 所示。从图 7 可以看出，两种输入尺寸的收敛情况几乎一致，并没有太大差别。但是训练所耗费的时间上两者差别较大，输入尺寸为 12×12 模型每秒会训练 3 000 个左右的样本，但是输入尺寸为 24×24 时模型每秒仅仅会训练 750 个左右的样本，主要原因就是在模型结构与参数完全一样的情况下，图片尺寸越小数据量越小，模型更容易进行计算。

由此可以得出，模型在有空间金字塔池化层的情况下，当输入图片的尺寸不同时，收敛情况几乎无差别，都会在训练到 8 000 次左右时收敛，即图片的输入尺寸不会影响模型的收敛情况。

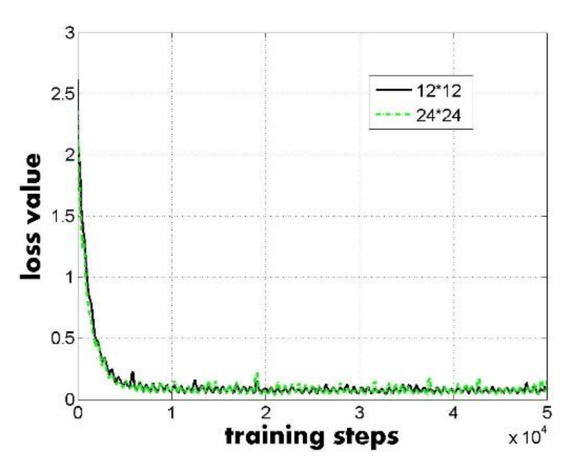


图 7　12×12 与 24×24 收敛情况对比

之后，在输入图片尺寸都是 24×24 的情况下，对加入金字塔池化层与不加金字塔池化层的模型进行了训练效果的对比，如图 8 所示。从图中可以看出，同样的训练数据，在加入金字塔池化方法以后，模型的收敛情况会相对更早。这里的主要原因是因为金字塔池化层是从粗到精的特征提取过程，可以更快速地筛选出从整体到局部的关键信息，对数据有着更好的泛化能力，所以添加了金字塔池化层会使卷积神经网络模型更早收敛。

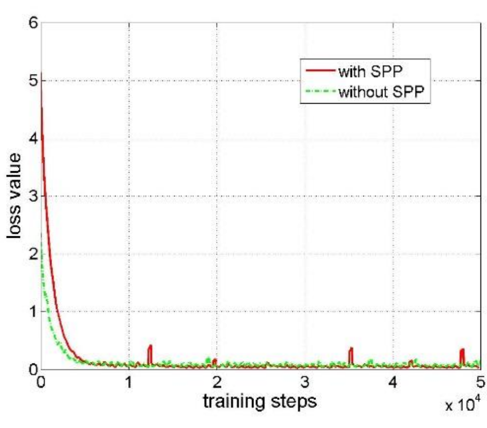


图 8　有无金字塔池化层训练对比

3.2　测试阶段实验结果分析

本文算法的测试阶段的测试数据集使用 FDDB 人脸数据集，主要测试了模型在这种数据集上的性能表现。FDDB 是全世界最具权威的人脸检测平台之一，包含 2 845 张图片，共有 5 171 个人脸作为测试集。测试集范围包括不同姿势、不同分辨率、旋转和遮挡等图片，同时还包括灰度图和彩色图，标准的人脸标注区域为椭圆形。这里在本文的测试中，采用了 FDDB 的另一种矩阵框标记形式，也就是通过标准椭圆标注信息进行转换，生成标准矩阵标注信息。

首先进行测试的是在 FDDB 数据集中，模型中添加了空间金字塔池化层(SPP)的检测性能与模型中没有空间金字塔池化层的检测性能的比较，如图 9 所示。图 9 所使用的 ROC 曲线又称感受型曲线，曲线上的各点反映着相同的感受性，曲线所覆盖面积越大说明效果越好。在 FDDB 数据集的结果检测中，统一用 ROC 曲线(receiver operating characteristic curve，受试者工作特征曲线)作为模型检测性能的标准。FDDB 数据集结果评测中会使用 continue score 和 discontinue score 来进行评价，由于两者都能反映出模型的性能表现，所以在测试中前两种测试结果曲线只展示了 continue score 的结果图，在最后与其他方法的对比时会进行两种结果图的展示。

在图 9 中可以看到，本文所采用的三级卷积神经网络在每级中都添加空间金字塔池化层时模型的检测效果是要相对较好的(这里输入模型的图片大小都是 24×24，区别就只有模型中是否加入了金字塔池化层)。另外，在图中添加空间金字塔池化层的网络模型的曲线可以更快地趋于平稳，这就说明模型的检测性能更加稳定，在样本量较少的情况下也能有较高的检测性能；而当样本的数量增加时，模型的效果又不会因为样本的增加而改变，也就表明金字塔池化层在提高模型的检测性能的同时，也提升了模型的鲁棒性。

接下来，使用了不同大小的输入图片对模型的性能进行测试，测试结果如图 10 所示。在测试中，分别将输入模型中的图片大小放缩成 12×12、18×18、24×24 三种尺寸，并进行三种输入大小的检测结果对比。在图 10 中可以明显看出，三种输入尺寸的检测结果曲线比较接近，其中输入尺寸为 24×24 时效果最好的，其次是 12×12，最后是 18×18。

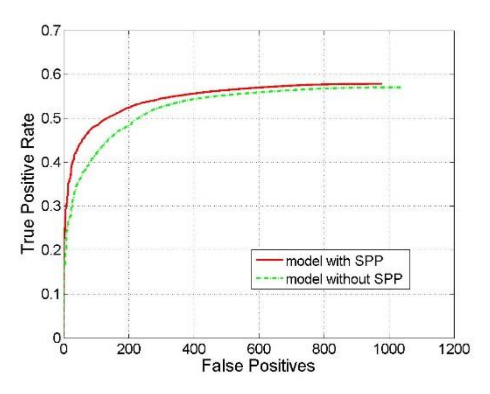


图 9　ROC 曲线-模型有无金字塔池化效果对比

在模型训练阶段中，由于模型只进行了 12×12 和 24×24 两种输入尺寸的交叉训练，所以在结果中可以发现输入尺寸为 12×12 以及 24×24 时的效果更好。而输入尺寸为 24×24 时的结果更好的原因有两点：a)图像大小放缩成 12×12 的过程中要损失更多的信息，这样模型提取到的特征相对于 24×24 大小的图片来说就会变少，使得检测性能降低；b)在图像输入之前需要进行金字塔图像的操作，当 12×12 的窗口扫描到完整的人脸区域时图像已经被放缩的很小，失真率很高，而 24×24 窗口扫描结果信息相对完整。因此，会出现 24×24 效果更好的情况。

虽然三种输入尺寸的检测结果有所不同，但是检测结果十分接近。另外，模型并没有进行 18×18 大小的图片的训练，检测结果依旧与其他两种比较相似。从这里就可以看出，加入金字塔池化层的级联卷积神经网络可以实现多种不同尺度大小的输入，并且模型在面对不同尺度的输入数据时可以有相近的检测准确率。

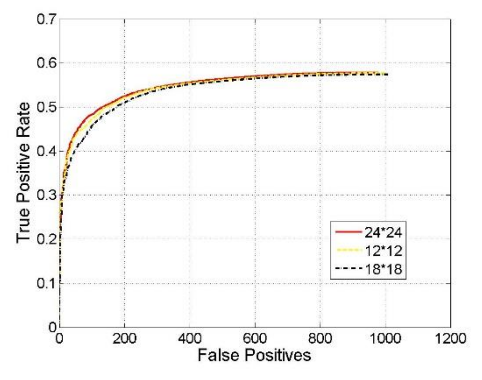
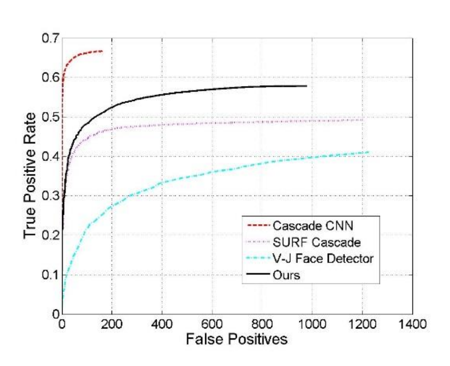


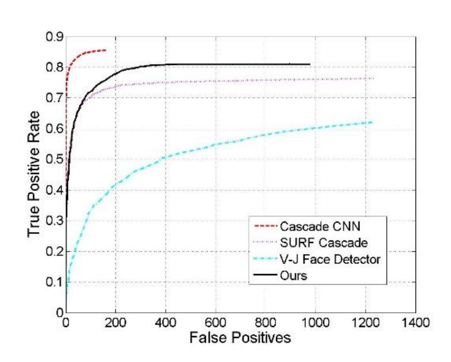
图 10　ROC 曲线-不同分辨率输入对比

最后，进行了本文所提出的模型与其他知名的性能对比，对比结果如图 11 所示。所对比的模型有经典的 V-J 人脸检测模型、基于级联 SURF 特征的人脸检测模型[16]，以及文献[10]所提出的级联卷积神经网络模型。可以发现，本文所提出的算法在 FDDB 数据集的测试中相对于 V-J 模型以及基于 SURF 特征的这些使用传统方法的人脸检测模型来说，显示出了更优秀的检测能力，并且算法对不同情况的人脸区域都有较为理想的检测结果。可见模型对于大数据集也具有不错的泛化能力。

另外，从图 11 还能发现，从本文所提出的加入金字塔池化的级联卷积神经网络与文献[10]中级联卷积神经网络的性能效果对比来看，本文所提出的算法性能相对较低，但是本文所提出的算法大幅度减少了模型的复杂性，也大大减少了训练模型的时间；同时，在进行检测时的所用时间也相对更少，文献[10]模型在使用一个 E5-2620 的 CPU 检测大小为 640×480 图片得到了大约 14 FPS 的检测速度，而本文的在 i7-4712MQ(性能要低)的 CPU 上检测同样大小的图片达到了大约 19 FPS 的检测速度(其他硬件也有影响)。最后，加入了金字塔池化层的卷积神经网络让模型的性能有所提升的同时，也使得模型可以实现对不同尺寸图片的检测。



(a)ROC 曲线-continue score



(b)ROC 曲线-discontinue score

图 11　本文所提出的模型与其他知名的性能对比结果

图 12 和 13 是本文模型的最终检测结果示意图。由于使用的滑动窗口的大小都是正方形，所以检测结果也是正方形的检测框。另外，从实际结果图来看，输入图片尺寸 24×24 与 12×12 的检测结果差异不大。整体来说，本文所提出的模型经过实际数据集测试后，在实现了卷积神经网络对图片的多尺度输入的同时，相对于传统算法在性能上也有提升。

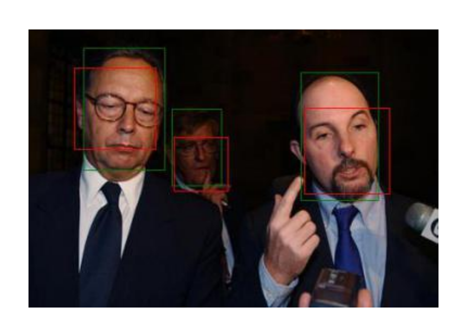


图 12　12×12 检测结果



图 13　24×24 检测结果

4　结束语

本文针对目前卷积神经网络不能支持多尺度数据输入以及单一结构模型对复杂情况泛化能力不强的问题，提出了基于金字塔池化的级联卷积神经网络的人脸检测算法。通过引入金字塔池化的方法，实现了卷积神经网络的多尺度输入，也在一定程度上对卷积神经网络的性能有所提升；然后结合了级联卷积神经网络的算法，使得模型不在局限于单一网络结构，使得模型在复杂情况下的泛化能力有所增强。

当然模型还有需要进一步完善的地方。接下来的工作将会首先将正方形检测框改成矩形甚至椭圆形，使得模型在人脸数据集中有更精确的检测精度；另外，在计算时间方面，将会进一步减少输入模型的候选框数量，进一步提高算法的运行效率。

参考文献：

[1] 卢宏涛，张秦川．深度卷积神经网络在计算机视觉中的应用研究综述[J]．数据采集与处理, 2016, 31 (1): 1-17.

[2] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436-444.

[3] Schmidhuber J. Deep learning in neural networks: an overview [J]. Neural Networks the Official Journal of the International Neural Networks Society, 2014, 61: 85-117.

[4] Guo Y, Oerlemans A, Lao S, et al. Deep learning for visual understanding [J]. Nerocomputing, 2016, 187 (C): 27-48.

[5] 孙志远，鲁成祥，史忠植，等．深度学习研究与发展 [J]．计算机科学, 2016, 43 (2): 1-8.

[6] 孙志军，薛磊，许阳明，等．深度学习研究综述 [J]．计算机应用研究, 2012, 29 (8): 2806-2810.

[7] Duc H H, Jung K. Applying tensorflow with convolutional neural networks to train data and recognize national flags [C]// Advanced Multimedia and Ubiquitous Engineering. 2017.

[8] Bianco S, Buzzelli M, Mazzini D, et al. Deep learning for logo recognition [J]. Neurocomputing, 2017, 245 (C): 23-30.

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014, 37 (9): 1904.

[10] Li Haoxiang, Lin Zhe, Shen Xiaohui, et al. A convolutional neural network cascade for face detection [C]// Proc of Computer Vision and Pattern Recognition. 2015: 5325-5334.

[11] Bouvrie J. Notes on convolutional neural networks [J]. Neural Nets, 2006, 31 (1): 1-17.

[12] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series [M]// The Handbook of Brain Theory and Neural Networks. 1995: 255-258.

[13] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. 2012: 1097-1105.

[14] Hao Biao, Kang D S. The research of face expression recognition based on CNN using tensorflow [J]. Journal of Advanced Information Technology and Convergence, 2017, 7.

[15] Viola P A, Jones M J. Rapid object detection using a boosted cascade of simple features [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition, 2003: 511-518.

[16] Li J, Wang T, Zhang Y. Face detection using SURF cascade [C]// Proc of IEEE International Conference on Computer Vision Workshops. 2012: 2183-2190.